(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 111522708 A
(43)申请公布日 2020.08.11

(21)申请号 202010280035.0

(22)申请日 2020.04.10

(71)申请人 星环信息科技(上海)有限公司
地址 200233 上海市徐汇区虹漕路88号B栋11-12楼

(72)发明人 荣国平 黄国成 顾胜晖

(74)专利代理机构 北京品源专利代理有限公司 11332
代理人 孟金喆

(51)Int.Cl.
*G06F 11/30*(2006.01)
*G06F 16/18*(2019.01)
*G06F 16/35*(2019.01)

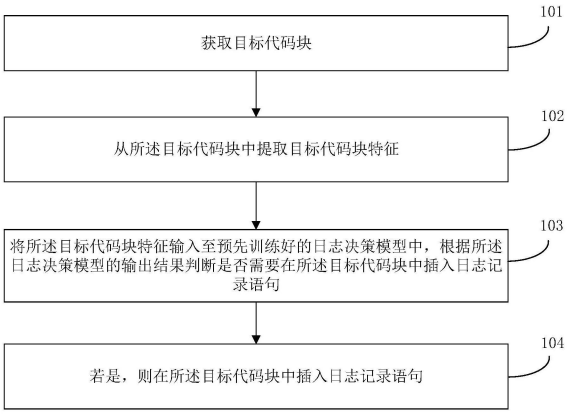
权利要求书3页 说明书16页 附图7页

(54)发明名称

一种日志记录方法、计算机设备及存储介质

(57)摘要

本发明实施例公开了一种日志记录方法、计算机设备及存储介质，其中，所述方法包括：获取目标代码块；从所述目标代码块中提取目标代码块特征；将所述目标代码块特征输入至预先训练好的日志决策模型中，根据所述日志决策模型的输出结果判断是否需要在所述目标代码块中插入日志记录语句；若是，则在所述目标代码块中插入日志记录语句。本发明实施例的技术方案能够可以合理、有效地进行日志记录，不仅能够减少琐碎无效的日志对系统资源的消耗，而且当系统异常时有助于软件开发人员通过合理有效的日志快速找到真正的问题。




CN 111522708 A

1.一种日志记录方法,其特征在于,包括:

获取目标代码块;

从所述目标代码块中提取目标代码块特征;

将所述目标代码块特征输入至预先训练好的日志决策模型中,根据所述日志决策模型的输出结果判断是否需要在所述目标代码块中插入日志记录语句;

若是,则在所述目标代码块中插入日志记录语句。

2.根据权利要求1所述的方法,其特征在于,从所述目标代码块中提取目标代码块特征,包括:

将所述目标代码块输入至源码分析工具中,根据所述源码分析工具的输出结果确定目标代码块特征。

3.根据权利要求1所述的方法,其特征在于,根据所述日志决策模型的输出结果判断是否需要在所述目标代码块中插入日志记录语句,包括:

获取所述日志决策模型的输出分数;

若所述输出分数大于预设分数阈值,则确定需要在所述目标代码块中插入日志记录语句;

若所述输出分数小于预设分数阈值,则确定不需要在所述目标代码块中插入日志记录语句。

4.根据权利要求1所述的方法,其特征在于,在获取目标代码块之前,还包括:

获取训练项目中的样本代码块;

根据所述样本代码块中是否包含日志记录语句对所述样本代码块进行样本标记;

提取样本标记后的样本代码块的样本代码块特征;

基于所述样本代码块特征对预设机器学习模型进行训练,生成日志决策模型。

5.根据权利要求4所述的方法,其特征在于,在基于所述样本代码块特征对预设机器学习模型进行训练之前,还包括:

对所述样本代码块特征中的文本特征依次进行驼峰转换,小写转换,删除停止词,词干提取和词根化处理以及频率-逆文档频率TF-IDF转换;

基于文本挖掘分类器对经过TF-IDF转换后的文本特征进行降维处理,生成数值文本特征。

6.根据权利要求5所述的方法,其特征在于,基于文本挖掘分类器对经过TF-IDF转换后的文本特征进行降维处理,生成数值文本特征,包括:

基于分层随机抽样将经过TF-IDF转换后的文本特征均分为第一样本和第二样本;

基于所述第一样本训练出第一文本挖掘分类器,并基于所述第二样本训练出第二文本挖掘分类器;

基于所述第一文本挖掘分类器为所述第一样本分配第一置信度分数,并基于所述第二文本挖掘分类器为所述第二样本分配第二置信度分数;

将所述第一置信度分数和所述第二置信度分数作为数值文本特征。

7.根据权利要求4所述的方法,其特征在于,在根据所述样本代码块中是否包含日志记录语句对所述样本代码块进行样本标记之后,还包括:

当所述样本代码块中包含日志记录语句时,确定所述样本代码块中包含的日志记录语

句的日志等级；

基于所述日志等级对所述样本代码块进行日志等级标记。

8.根据权利要求7所述的方法，其特征在于，所述日志等级包括致命等级、错误等级、警告等级、信息等级、调试等级和跟踪等级。

9.根据权利要求4所述的方法，其特征在于，所述预设机器学习模型包括随机森林算法、朴素贝叶斯算法、自适应增强Adaboost算法、支持向量机算法。

10.根据权利要求1-9任一所述的方法，其特征在于，代码块特征包括文本特征和句法特征；

其中，所述文本特征包括代码块的结构特征、代码块调用的方法名称、代码块中声明的变量名称、代码块的类型以及触发策略类型；所述代码块的结构特征包括代码块的源代码行SLOC、代码块调用的方法数目、代码块中声明的变量数目以及代码块中包含日志；

所述句法特征包括代码块中是否存在throw语句、是否存在assert语句、是否存在返回值以及是否存在中断异常类型中的任一种。

11.一种计算机设备，包括处理器和存储器，所述存储器用于存储指令，当所述指令执行时使得所述处理器执行以下操作：

获取目标代码块；

从所述目标代码块中提取目标代码块特征；

将所述目标代码块特征输入至预先训练好的日志决策模型中，根据所述日志决策模型的输出结果判断是否需要在所述目标代码块中插入日志记录语句；

若是，则在所述目标代码块中插入日志记录语句。

12.根据权利要求11所述的计算机设备，其特征在于，所述处理器是设置为通过以下方式从所述目标代码块中提取目标代码块特征：

将所述目标代码块输入至源码分析工具中，根据所述源码分析工具的输出结果确定目标代码块特征。

13.根据权利要求11所述的计算机设备，其特征在于，所述处理器是设置为通过以下方式根据所述日志决策模型的输出结果判断是否需要在所述目标代码块中插入日志记录语句：

获取所述日志决策模型的输出分数；

若所述输出分数大于预设分数阈值，则确定需要在所述目标代码块中插入日志记录语句；

若所述输出分数小于预设分数阈值，则确定不需要在所述目标代码块中插入日志记录语句。

14.根据权利要求11所述的计算机设备，其特征在于，在获取目标代码块之前，所述处理器还设置为：

获取训练项目中的样本代码块；

根据所述样本代码块中是否包含日志记录语句对所述样本代码块进行样本标记；

提取样本标记后的样本代码块的样本代码块特征；

基于所述样本代码块特征对预设机器学习模型进行训练，生成日志决策模型。

15.根据权利要求14所述的计算机设备，其特征在于，在基于所述样本代码块特征对预

设机器学习模型进行训练之前，所述处理器还设置为：

对所述样本代码块特征中的文本特征依次进行驼峰转换，小写转换，删除停止词，词干提取和词根化处理以及频率-逆文档频率TF-IDF转换；

基于文本挖掘分类器对经过TF-IDF转换后的文本特征进行降维处理，生成数值文本特征。

16.根据权利要求15所述的计算机设备，其特征在于，所述处理器是设置为通过以下方式基于文本挖掘分类器对经过TF-IDF转换后的文本特征进行降维处理，生成数值文本特征，包括：

基于分层随机抽样将经过TF-IDF转换后的文本特征均分为第一样本和第二样本；

基于所述第一样本训练出第一文本挖掘分类器，并基于所述第二样本训练出第二文本挖掘分类器；

基于所述第一文本挖掘分类器为所述第一样本分配第一置信度分数，并基于所述第二文本挖掘分类器为所述第二样本分配第二置信度分数；

将所述第一置信度分数和所述第二置信度分数作为数值文本特征。

17.根据权利要求14所述的计算机设备，其特征在于，在根据所述样本代码块中是否包含日志记录语句对所述样本代码块进行样本标记之后，所述处理器还设置为：

当所述样本代码块中包含日志记录语句时，确定所述样本代码块中包含的日志记录语句的日志等级；

基于所述日志等级对所述样本代码块进行日志等级标记。

18.根据权利要求17所述的计算机设备，其特征在于，所述日志等级包括致命等级、错误等级、警告等级、信息等级、调试等级和跟踪等级。

19.根据权利要求14所述的计算机设备，其特征在于所述预设机器学习模型包括随机森林算法、朴素贝叶斯算法、自适应增强Adaboost算法、支持向量机算法。

20.根据权利要求11-19任一所述的计算机设备，其特征在于，代码块特征包括文本特征和句法特征；

其中，所述文本特征包括代码块的结构特征、代码块调用的方法名称、代码块中声明的变量名称、代码块的类型以及触发策略类型；所述代码块的结构特征包括代码块的源代码行SLOC、代码块调用的方法数目、代码块中声明的变量数目以及代码块中包含日志；

所述句法特征包括代码块中是否存在throw语句、是否存在assert语句、是否存在返回值以及是否存在中断异常类型中的任一种。

21.一种计算机存储介质，其上存储有计算机程序，其特征在于，该程序被处理器执行时实现如权利要求1-10中任一所述的日志记录方法。

# 一种日志记录方法、计算机设备及存储介质

## 技术领域

[0001] 本发明实施例涉及数据处理技术领域，尤其涉及一种日志记录方法、计算机设备及存储介质。

## 背景技术

[0002] 随着DevOps的提出与发展，日志的记录与分析愈发受到人们的重视。但是当下，人们更加重视如何从得到的日志中分析问题或获取信息，却相对忽视日志的记录。在相关的调查中，在15797个github上最受关注的开源java项目中，有60％从不记录日志，而记录日志的项目中，有60.2％的日志语句中不包含任何参数信息(这意味着日志不能有效的展示系统状态、提供错误信息，是无效的日志)。日志的记录是分析日志的基础，只有在日志被高效、合理地记录的前提下，日志分析的相关技术或研究才具有意义。

[0003] 日志记录是收集系统运行信息以用于事后分析的一种具有实际意义的通用编程实践。例如，Web服务器应用程序可以记录时间戳，客户端IP地址和完整URL，包括未找到文件的异常情况下所请求的文件的名称。在调试时，软件开发人员可以使用该执行信息。所以日志记录已经成为将软件系统的关键运行信息(例如，状态、事件)记录到日志中以便事后分析的主要方式。日志记录通常通过在源代码中插入日志语句(例如，printf()、log.warn())来实现。

[0004] 然而，现有技术中日志记录的内容是海量的，然而日志记录过多将会产生一系列问题：首先，日志意味着更多的代码，需要花费时间来编写和维护；其次，日志记录消耗额外的系统资源(例如，CPU和I/O)，并且对系统操作具有显著的性能影响；另外，过多的日志记录会产生许多琐碎而无用的日志，最终掩盖真正重要的信息，从而使得当系统异常时软件开发人员很难找到真正的问题。因此，合理、有效的日志记录变得至关重要。

## 发明内容

[0005] 本发明实施例提供一种日志记录方法、计算机设备及存储介质，以合理、有效地进行日志记录。

[0006] 第一方面，本发明实施例提供了一种日志记录方法，包括：

[0007] 获取目标代码块；

[0008] 从所述目标代码块中提取目标代码块特征；

[0009] 将所述目标代码块特征输入至预先训练好的日志决策模型中，根据所述日志决策模型的输出结果判断是否需要在所述目标代码块中插入日志记录语句；

[0010] 若是，则在所述目标代码块中插入日志记录语句。

[0011] 第二方面，本发明实施例还提供了一种日志记录装置，包括：

[0012] 目标代码块获取模块，用于获取目标代码块；

[0013] 目标代码块特征提取模块，用于从所述目标代码块中提取目标代码块特征；

[0014] 日志记录语句判断模块，用于将所述目标代码块特征输入至预先训练好的日志决

策模型中，根据所述日志决策模型的输出结果判断是否需要在所述目标代码块中插入日志记录语句；

[0015] 日志记录语句插入模块，用于若根据所述日志决策模型的输出结果判断出需要在所述目标代码块中插入日志记录语句，则在所述目标代码块中插入日志记录语句。

[0016] 第三方面，本发明实施例还提供了一种计算机设备，包括处理器和存储器，存储器用于存储指令，当指令执行时使得处理器执行以下操作：

[0017] 获取目标代码块；

[0018] 从所述目标代码块中提取目标代码块特征；

[0019] 将所述目标代码块特征输入至预先训练好的日志决策模型中，根据所述日志决策模型的输出结果判断是否需要在所述目标代码块中插入日志记录语句；

[0020] 若是，则在所述目标代码块中插入日志记录语句。

[0021] 第四方面，本发明实施例还提供了一种存储介质，存储介质用于存储指令，指令用于执行：

[0022] 获取目标代码块；

[0023] 从所述目标代码块中提取目标代码块特征；

[0024] 将所述目标代码块特征输入至预先训练好的日志决策模型中，根据所述日志决策模型的输出结果判断是否需要在所述目标代码块中插入日志记录语句；

[0025] 若是，则在所述目标代码块中插入日志记录语句。

[0026] 本发明实施例通过获取目标代码块，并从目标代码块中提取目标代码块特征，然后将目标代码块特征输入至预先训练好的日志决策模型中，根据日志决策模型的输出结果判断是否需要在目标代码块中插入日志记录语句；若是，则在目标代码块中插入日志记录语句，可以合理、有效地进行日志记录，减少日志记录的决策时间，不仅能够减少琐碎无效的日志对系统资源的消耗，而且当系统异常时有助于软件开发人员通过合理有效的日志快速找到真正的问题，提高开发效率。

## 附图说明

[0027] 图1是本发明实施例一提供的一种日志记录方法的流程图；

[0028] 图2是本发明实施例二提供的一种日志记录方法的流程图；

[0029] 图3是本发明实施例二提供的一种日志记录决策的过程示意图；

[0030] 图4是本发明实施例二提供的数值文本特征的生成过程示意图；

[0031] 图5是本发明实施例三提供的一种日志记录方法的流程图；

[0032] 图6是本发明实施例四提供的一种日志记录装置的示意图；

[0033] 图7是本发明实施例五提供的基于机器学习的日志决策推荐插件的结构图；

[0034] 图8为本发明实施例六提供的一种计算机设备的结构示意图。

## 具体实施方式

[0035] 下面结合附图和实施例对本发明作进一步的详细说明。可以理解的是，此处所描述的具体实施例仅仅用于解释本发明，而非对本发明的限定。

[0036] 另外还需要说明的是，为了便于描述，附图中仅示出了与本发明相关的部分而非

全部内容。在更加详细地讨论示例性实施例之前应当提到的是，一些示例性实施例被描述成作为流程图描绘的处理或方法。虽然流程图将各项操作(或步骤)描述成顺序的处理，但是其中的许多操作可以被并行地、并发地或者同时实施。此外，各项操作的顺序可以被重新安排。当其操作完成时所述处理可以被终止，但是还可以具有未包括在附图中的附加步骤。所述处理可以对应于方法、函数、规程、子例程、子程序等等。

[0037] 本文使用的术语“目标代码块”即为待分析是否需要插入日志记录语句的代码块。

[0038] 本文使用的术语“目标代码块特征”可以是提取的目标代码块的特征，如目标代码块调用的方法、名称等。

[0039] 本文使用的术语“日志决策模型”可以是用于快速判断是否需要在目标代码块中插入日志记录语句的机器学习模型。

[0040] 本文使用的术语“样本代码块”可以为用于训练日志决策模型的样本数据。

[0041] 本文使用的术语“样本代码块特征”可以是提取的样本代码块的特征，如样本代码块调用的方法、名称等。

[0042] 本文使用的术语“日志等级”可以是用于描述记录日志中所记录信息的详细程度。

[0043] 为了便于理解，将本发明实施例的主要发明构思进行简述。

[0044] 实施例一

[0045] 图1是本发明实施例一提供的一种日志记录方法的流程图，本实施例可适用于对代码块进行日志记录决策的情况，该方法可以由日志记录装置来执行，该装置可以由软件和/或硬件的方式来实现，并一般可集成在计算机设备中。相应的，如图1所示，该方法包括如下操作：

[0046] 步骤101、获取目标代码块。

[0047] 其中，目标代码块可以是待分析是否需要插入日志记录语句的代码块。日志记录语句用于记录信息或者处理一些系统运行时遇到的错误，在触发执行后可以形成记录日志。可选的，在开发新项目时，获取新项目中待分析的是否需要插入日志记录语句的代码块作为目标代码块。

[0048] 步骤102、从所述目标代码块中提取目标代码块特征。

[0049] 在本发明实施例中，可以通过预设特征模型，对目标代码块进行特征提取，得到目标代码块特征。可选的，从所述目标代码块中提取目标代码块特征，包括：将所述目标代码块输入至源码分析工具中，根据所述源码分析工具的输出结果确定目标代码块特征。其中，源码分析工具可以为JavaParser工具，JavaParser工具不仅可以提取代码块特征，还可以分析源码。

[0050] 其中，目标代码块特征描述的是目标代码块的特征。可选的，目标代码块特征包括文本特征和句法特征；其中，所述文本特征包括代码块的结构特征、代码块调用的方法名称、代码块中声明的变量名称、代码块的类型以及触发策略类型；所述代码块的结构特征包括代码块的源代码行SLOC、代码块调用的方法数目、代码块中声明的变量数目以及代码块中包含日志；所述句法特征包括代码块中是否存在throw语句、是否存在assert语句、是否存在返回值以及是否存在中断异常类型中的任一种。

[0051] 在本发明实施例中，可以提取目标代码块的46个特征作为目标代码块特征。具体的，46个目标代码块特征可以如下表1所示：

[0052] 表1

[0053]

| 标号 | 特征 | 域 | 类型 | 描述 |
|---|---|---|---|---|
| 1 | Size of Try Block | Try/Catch | Numeric | 代码块的代码行数，代码行数越多，意味着需要日志的可能性越大 |
| 2 | Size of Method BT | Method BT | Numeric | |
| 3 | Catch Exception Type | Try/Catch | Textual | 异常类型是影响日志决策的一个重要方面 |
| 4 | Previous Catch Blocks | Try/Catch | Boolean | Catch 块前有无 Catch 块，catch 块的数量说明了，可出现的异常几率越大 |
| 5 | Logged Previous Catch Blocks | Try/Catch | Boolean | Catch 块前的 catch 有无日志语句，已有日志语句是影响新 catch 块的日志决策的一个方面 |
| 6 | Logged Try Block | Try/Catch | Boolean | Try 后的 catch 有无日志语句，try 块的日志语句会影响 catch 块的日志决策 |
| 7 | Logged Method BT | Method BT | Boolean | |
| 8 | LogCountTry Block | Try/Catch | Numeric | Try 块中的日志语句数量，try 块的日志语句会数量影响 catch 块的日志决策 |
| 9 | LogCountin Method BT | Method BT | Numeric | |
| 10 | Log Levels in TryBlock | Try/Catch | Textual | 前面 try 块中的日志语句级别，try 块的日志语句等级会数量影响 catch 块的日志决策 |
| 11 | LogLeelsin Method BT | Method BT | Textual | |
| 12 | Operators in Try Block | Try/Catch | Textual | |
| 13 | Operators in Method BT | Method BT | Textual | |
| 14 | Count of Operators in Try Block | Try/Catch | Numeric | |
| 15 | Count of Operators in Method BT | Method BT | Numeric | |
| 16 | Variable Count in Try Block | Try/Catch | Numeric | 代码块中声明变量的个数，变量个数越多，可能需要日志的概率越大； |
| 17 | Variable Count in Method BT | Method BT | Numeric | |
| 18 | Method Call Count in Try Block | Try/Catch | Numeric | 代码块中函数调用的个数，函数调用个数越多，可能需要日志的概率越大； |
| 19 | Method Call Count in Method BT | Method BT | Numeric | |
| 20 | Container Method have Parameter | Other | Boolean | 包含此块的函数有无参数，这是影响日志决策的 |
| 21 | Container Method | Other | Numeric | |

[0054]

[0055]

| | | | | |
|---|---|---|---|---|
| | Parameter Count | | | 一个方面 |
| 22 | Container Method Parameters (Type) | Contextual | Textual | 包含此块的函数的参数类型，这是影响日志决策的一个方面 |
| 23 | Container Method Parameters(Name) | Other | Textual | |
| 24 | IF in Try | Try/Catch | Boolean | Try 中有无 if 语句，这是影响日志决策的一个方面<br>If 语句块的个数，if 判断越多，需要日志语句的概率越大 |
| 25 | IF in Method BT | Method BT | Boolean | |
| 26 | IF Count in Try Block | Try/Catch | Numeric | |
| 27 | IFCount in Method BT | Method BT | Numeric | |
| 28 | Container Package Name | Other | Textual | 开发人员会对在开发当中会给变量、方法、包、类、类型有意义的名称。这些都会对日志决策有影响 |
| 29 | Container Class Name | Other | Textual | |
| 30 | Container Method Name | Other | Textual | |
| 31 | Variable Name in Try Block | Try/Catch | Textual | 开发人员会对在开发当中会给变量、方法、包、类、类型有意义的名称。这些都会对日志决策有影响 |
| 32 | Variable Name in Method BT | Method BT | Textual | |
| 33 | Method Call Name in Try Block | Try/Catch | Textual | 开发人员会对在开发当中会给变量、方法、包、类、类型有意义的名称。这些都会对日志决策有影响 |
| 34 | Method Call Name in Method BT | Method BT | Textual | |
| 35 | Throw/Throws in Try Block | Try/Catch | Boolean | 有无 throw，这是影响日志决策的一个方面 |
| 36 | Throw/Throws in Catch Block | Try/Catch | Boolean | |
| 37 | Throw/Throws in Method BT | Method BT | Boolean | |
| 38 | Return in Try Block | Try/Catch | Boolean | 有无 return，这是影响日志决策的一个方面 |
| 39 | Return in Catch Block | Try/Catch | Boolean | |
| 40 | Return in Method BT | Method BT | Boolean | |
| 41 | Assert in Try Block | Try/Catch | Boolean | 有无 assert 语句，这是影响日志决策的一个方面 |
| 42 | Assert in Catch Block | Try/Catch | Boolean | |
| 43 | Assert in Method BT | Method BT | Boolean | |

[0056]

| | | | | |
|---|---|---|---|---|
| 44 | Thread.Sleep in Try Block | Try/Catch | Boolean | 有无Thread.Sleep().，这是影响日志决策的一个方面 |
| 45 | Interrupted Exception Type | Try/Catch | Boolean | 有无中断异常类型，这是影响日志决策的一个方面 |
| 46 | Exception Object ‘Ignore’ in Catch | Try/Catch | Boolean | 异常对象名称是不是ignore，这是影响日志决策的一个方面 |

[0057] 步骤103、将所述目标代码块特征输入至预先训练好的日志决策模型中，根据所述日志决策模型的输出结果判断是否需要在所述目标代码块中插入日志记录语句。

[0058] 在本发明实施例中，日志决策模型可以理解为在输入目标代码块特征后快速确定是否需要在目标代码块中插入日志记录语句的学习模型。日志决策模型可以是基于包含有日志记录语句的代码块的特征规律生成的网络模型，即将包含日志记录语句和不包含日志记录语句的代码块特征作为训练样本，根据预设的机器学习模型对该训练样本进行训练、学习，生成日志决策模型。

[0059] 可选的，根据所述日志决策模型的输出结果判断是否需要在所述目标代码块中插入日志记录语句，包括：获取所述日志决策模型的输出分数；若所述输出分数大于预设分数阈值，则确定需要在所述目标代码块中插入日志记录语句；若所述输出分数小于预设分数阈值，则确定不需要在所述目标代码块中插入日志记录语句。

[0060] 在本发明实施例中，将目标代码块特征输入至日志决策模后，日志决策模型的输出结果为置信度分数，该置信度分数反映了日志决策模型对目标代码块特征进行分析后，预测出在目标代码块中插入日志记录语句的概率大小(也即可能性大小)。当该置信度分数(日志决策模型的输出分数)大于至预设分数阈值时，确定需要在目标代码块中插入日志记录语句；当该置信度分数小于预设分数阈值时，确定不需要在目标代码块中插入日志记录语句。

[0061] 步骤104、若是，则在所述目标代码块中插入日志记录语句。

[0062] 在本发明实施例中，当根据日志决策模型的输出结果判断出需要在目标代码块中插入日志记录语句时，在目标代码块中插入日志记录语句，以进行适当的信息记录，从而生成记录日志。

[0063] 本发明实施例通过获取目标代码块，并从目标代码块中提取目标代码块特征，然后将目标代码块特征输入至预先训练好的日志决策模型中，根据日志决策模型的输出结果判断是否需要在目标代码块中插入日志记录语句；若是，则在目标代码块中插入日志记录语句，可以合理、有效地进行日志记录，减少日志记录的决策时间，不仅能够减少琐碎无效的日志对系统资源的消耗，而且当系统异常时有助于软件开发人员通过合理有效的日志快速找到真正的问题，提高开发效率。

[0064] 在本发明的一个可选实施例中，在获取目标代码块之前，还包括：获取训练项目中的样本代码块；根据所述样本代码块中是否包含日志记录语句对所述样本代码块进行样本标记；提取样本标记后的样本代码块的样本代码块特征；基于所述样本代码块特征对预设

机器学习模型进行训练,生成日志决策模型。

[0065] 示例性的,可以选取现有优秀项目作为训练项目,例如,可以将GitHub上的star排名靠前的几个项目选取为训练项目,如可将Tomcat,CloudStack和Hadoop这三个项目中的任一一个或多个作为训练项目。由于这些项目都是由Apache软件基金会之类行业标杆维护的项目,因此,可以有效保证训练日志决策模型时样本数据集的可靠性。在本发明实施例中,可以基于Javaparser工具从训练项目中获取代码块,并作为样本代码块。判断样本代码块中是否包含至少一个日志记录语句,当样本代码块中包含至少一个日志记录语句时,可以将该样本代码块标记为“logged”;当样本代码块中不包含任何日志记录语句时,将该样本代码块标记为“unlogged”。其中,在判断样本代码块中是否包含日志记录语句时,可以通过多个正则表达式进行字符串匹配,以检测样本代码块中是否包含日志记录语句。

[0066] 示例性的,对样本代码块进行样本标记后,提取样本标记后的样本代码块的特征,作为样本代码块特征。其中,样本代码块特征描述的是样本代码块的特征。可选的,可以将样本代码块输入至源码分析工具中,根据源码分析工具的输出结果确定样本代码块特征。可选的,样本代码块特征包括文本特征和句法特征;其中,所述文本特征包括代码块的结构特征、代码块调用的方法名称、代码块中声明的变量名称、代码块的类型以及触发策略类型;所述代码块的结构特征包括代码块的源代码行SLOC、代码块调用的方法数目、代码块中声明的变量数目以及代码块中包含日志;所述句法特征包括代码块中是否存在throw语句、是否存在assert语句、是否存在返回值以及是否存在中断异常类型中的任一种。

[0067] 在本发明实施例中,可以提取样本代码块的46个特征作为样本代码块特征。具体的,46个目标代码块特征可以如上述表1所示。其中,按照数据类型可以将样本代码块特征特征可以分为数值特征、布尔特征、文本特征,例如代码块行数为数值特征,是否有断言语句为布尔特征,异常类型为文本特征。

[0068] 在本发明的一个可选实施例中,在基于所述样本代码块特征对预设机器学习模型进行训练之前,还包括:对所述样本代码块特征中的文本特征依次进行驼峰转换,小写转换,删除停止词,词干提取和词根化处理以及频率-逆文档频率TF-IDF转换;基于文本挖掘分类器对经过TF-IDF转换后的文本特征进行降维处理,生成数值文本特征。这样设置的好处在于,可以有效去除样本代码块特征中的文本特征的冗余信息,并将文本特征转换为数值文本特征。

[0069] 示例性的,可以首先使用驼峰转换技术将样本代码块特征中的文本特征中连接的术语分开。例如,'LoginFailure'将转换为'Login'和'Failure'。其次,对驼峰转换后的文本特征进行小写转换处理,将样本代码块特征中的文本特征中的术语转换为小写。例如,将术语'Login'将转换为'login'。停用词是文档中经常出现的术语,因为,它们出现在大多数的文档中,被认为是非内容承载的特征。因此,对小写转换处理后的样本代码块特征中的文本特征继续进行删除停用词处理,从文本中删除英语停用词,例如'the','is'这些词是对于文本分类毫无意义的词。然后,继续对删除停用词后的样本代码块特征中的文本特征进行词干提取和词根化处理,以基于词干将文件特征中的术语转换为其词根形式,其中,词干是将变形词减少到其根形式的过程。例如,术语“modifier”和“modify”将转换为根词“modifi”。而词根形式的转换可以降低文本特征的维度,有利于减少特征空间的时间和空间复杂性。最后,将所有项转换为它们的TF-IDF(频率-逆文档频率)表示,其中,TF-IDF是一

种数值统计，用于识别样本代码块中文本特征中单词的重要性。

[0070] 在对样本代码块特征中的文本特征进行TF-IDF转换会生成数千个特征，若直接基于所有这些TF-IDF转换后的文本特征直接与数字特征以及布尔特征对日志决策模型进行训练，会稀释数值特征和布尔特征的权重，影响日志决策模型进行是否需要插入日志记录语句判断的准确性。

[0071] 在本发明的一个可选实施例中，基于文本挖掘分类器对经过TF-IDF转换后的文本特征进行降维处理，生成数值文本特征，包括：基于分层随机抽样将经过TF-IDF转换后的文本特征均分为第一样本和第二样本；基于所述第一样本训练出第一文本挖掘分类器，并基于所述第二样本训练出第二文本挖掘分类器；基于所述第一文本挖掘分类器为所述第一样本分配第一置信度分数，并基于所述第二文本挖掘分类器为所述第二样本分配第二置信度分数；将所述第一置信度分数和所述第二置信度分数作为数值文本特征。

[0072] 示例性的，使用分层随机抽样将经过TF-IDF处理的文本特征均分成第一样本和第二样本；然后可以根据朴素贝叶斯算法分别学习出与第一样本对应的第一文本挖掘分类器，以及与第二样本对应的第二文本挖掘分类器；利用第一文本挖掘分类器为第二样本分配第二置信度分数，以及利用第二文本挖掘分类器为第一样本分配第一置信度分数；第一置信度分数和第二置信度分数即为数值文本特征。

[0073] 通过上述方案，可以将上述46个样本代码块特征转换为19个特征，其中，包括11个布尔特征，7个数值特征和1个数值文本特征。然后，将上述19个特征输入至预设的机器学习模型进行训练，生成日志决策模型。其中，所述预设机器学习模型包括随机森林算法、朴素贝叶斯算法、自适应增强Adaboost算法、支持向量机算法。

[0074] 在本发明的一个可选实施例中，在根据所述样本代码块中是否包含日志记录语句对所述样本代码块进行样本标记之后，还包括：当所述样本代码块中包含日志记录语句时，确定所述样本代码块中包含的日志记录语句的日志等级；基于所述日志等级对所述样本代码块进行日志等级标记。这样设置的好处在于，日志决策模型不仅可以预测出是否需要在目标代码块中插入日志记录语句，而且当确定需要在目标代码块中插入日志记录语句时，还可以准确预测出待插入日志记录语句的日志等级。

[0075] 其中，日志等级用于描述记录日志中所记录信息的详细程度。可选的，日志等级包括致命等级、错误等级、警告等级、信息等级、调试等级和跟踪等级。其中，致命等级、错误等级、警告等级、信息等级、调试等级和跟踪等级的重要性等级逐级递减，而且重要性等级越低，记录日志中所记录信息越详细。其中，致命级别是指非常严重的错误事件，可能导致应用程序中止。错误级别是指仍然允许应用程序继续运行的错误事件。警告级别是指潜在的有害情况。信息级别是指信息性消息，这些消息在粗粒度级别高度突出应用程序的进度。调试级别是指对调试应用程序最有用的精细信息事件。跟踪级别是指比“调试”更细粒度的信息事件。

[0076] 在本发明实施例中，当样本代码块中包含日志记录语句时，进一步判断样本代码块中包含的日志记录语句的日志等级，并根据日志等级对相应的样本代码快进行日志等级标记。其中，日志等级不同，日志等级标记不同，例如，可以用不同的数字标记不同的日志等级，其中，日志等级的重要性等级越高，进行日志等级标记的数字越小。需要说明的是，本发明实施例对日志等级的标记方式不做限定。

[0077] 实施例二

[0078] 图2是本发明实施例二提供的一种日志记录方法的流程图，本实施例以上述实施例为基础进行具体化，在本实施例中，在获取目标代码块之前，还包括：获取训练项目中的样本代码块；根据所述样本代码块中是否包含日志记录语句对所述样本代码块进行样本标记；提取样本标记后的样本代码块的样本代码块特征；基于所述样本代码块特征对预设机器学习模型进行训练，生成日志决策模型。相应的，如图2所示，本实施例的方法可以包括：

[0079] 步骤201、获取训练项目中的样本代码块。

[0080] 步骤202、根据所述样本代码块中是否包含日志记录语句对所述样本代码块进行样本标记。

[0081] 步骤203、提取样本标记后的样本代码块的样本代码块特征。

[0082] 可选的，将样本标记后的样本代码块输入至源码分析工具中，根据源码分析工具的输出结果确定样本代码块特征。可选的，样本代码块特征包括文本特征和句法特征；其中，所述文本特征包括代码块的结构特征、代码块调用的方法名称、代码块中声明的变量名称、代码块的类型以及触发策略类型；所述代码块的结构特征包括代码块的源代码行SLOC、代码块调用的方法数目、代码块中声明的变量数目以及代码块中包含日志；所述句法特征包括代码块中是否存在throw语句、是否存在assert语句、是否存在返回值以及是否存在中断异常类型中的任一种。

[0083] 步骤204、基于所述样本代码块特征对预设机器学习模型进行训练，生成日志决策模型。

[0084] 可选的，所述预设机器学习模型包括随机森林算法、朴素贝叶斯算法、自适应增强Adaboost算法、支持向量机算法。

[0085] 可选的，在基于所述样本代码块特征对预设机器学习模型进行训练之前，还包括：对所述样本代码块特征中的文本特征依次进行驼峰转换，小写转换，删除停止词，词干提取和词根化处理以及频率-逆文档频率TF-IDF转换；基于文本挖掘分类器对经过TF-IDF转换后的文本特征进行降维处理，生成数值文本特征。

[0086] 在本发明实施例中，图3为本发明实施例二提供的日志记录决策的过程示意图。如图3所示，由于文本特征无法直接作为机器学习模型的输入数据进行学习，因此需要对其进行一系列预处理，除去冗余信息并转化为数字表示。但是文本特征预处理中生成的特征维度过大，会稀释掉数字特征和布尔特征在模型中的作用，因此，需要再利用文本挖掘器对预处理生成的特征进行降维处理，生成数值文本特征。

[0087] 可选的，基于文本挖掘分类器对经过TF-IDF转换后的文本特征进行降维处理，生成数值文本特征，包括：基于分层随机抽样将经过TF-IDF转换后的文本特征均分为第一样本和第二样本；基于所述第一样本训练出第一文本挖掘分类器，并基于所述第二样本训练出第二文本挖掘分类器；基于所述第一文本挖掘分类器为所述第一样本分配第一置信度分数，并基于所述第二文本挖掘分类器为所述第二样本分配第二置信度分数；将所述第一置信度分数和所述第二置信度分数作为数值文本特征。如图4所示，为本发明实施例二提供的数值文本特征的生成过程示意图。

[0088] 步骤205、获取目标代码块。

[0089] 步骤206、从所述目标代码块中提取目标代码块特征。

[0090] 可选的，将目标代码块输入至源码分析工具中，根据源码分析工具的输出结果确定目标代码块特征。可选的，目标代码块特征包括文本特征和句法特征；其中，所述文本特征包括代码块的结构特征、代码块调用的方法名称、代码块中声明的变量名称、代码块的类型以及触发策略类型；所述代码块的结构特征包括代码块的源代码行SLOC、代码块调用的方法数目、代码块中声明的变量数目以及代码块中包含日志；所述句法特征包括代码块中是否存在throw语句、是否存在assert语句、是否存在返回值以及是否存在中断异常类型中的任一种。

[0091] 步骤207、将所述目标代码块特征输入至所述预先训练好的日志决策模型中，根据所述日志决策模型的输出结果判断是否需要在所述目标代码块中插入日志记录语句。

[0092] 可选的，根据所述日志决策模型的输出结果判断是否需要在所述目标代码块中插入日志记录语句，包括：获取所述日志决策模型的输出分数；若所述输出分数大于预设分数阈值，则确定需要在所述目标代码块中插入日志记录语句；若所述输出分数小于预设分数阈值，则确定不需要在所述目标代码块中插入日志记录语句。

[0093] 步骤208、若根据所述日志决策模型的输出结果判断出需要在所述目标代码块中插入日志记录语句，则在所述目标代码块中插入日志记录语句。

[0094] 本发明实施例通过获取目标代码块，并从目标代码块中提取目标代码块特征，然后将目标代码块特征输入至预先训练好的日志决策模型中，根据日志决策模型的输出结果判断是否需要在目标代码块中插入日志记录语句；若是，则在目标代码块中插入日志记录语句，可以合理、有效地进行日志记录，减少日志记录的决策时间，不仅能够减少琐碎无效的日志对系统资源的消耗，而且当系统异常时有助于软件开发人员通过合理有效的日志快速找到真正的问题，提高开发效率。

[0095] 实施例三

[0096] 图5是本发明实施例三提供的一种日志记录方法的流程图，本实施例以上述实施例为基础进行具体化，在根据所述样本代码块中是否包含日志记录语句对所述样本代码块进行样本标记之后，还包括：当所述样本代码块中包含日志记录语句时，确定所述样本代码块中包含的日志记录语句的日志等级；基于所述日志等级对所述样本代码块进行日志等级标记。相应的，如图5所示，本实施例的方法可以包括：

[0097] 步骤501、获取训练项目中的样本代码块。

[0098] 步骤502、根据所述样本代码块中是否包含日志记录语句对所述样本代码块进行样本标记。

[0099] 步骤503、当所述样本代码块中包含日志记录语句时，确定所述样本代码块中包含的日志记录语句的日志等级。

[0100] 可选的，所述日志等级包括致命等级、错误等级、警告等级、信息等级、调试等级和跟踪等级。

[0101] 步骤504、基于所述日志等级对所述样本代码块进行日志等级标记。

[0102] 步骤505、提取样本标记和日志等级标记后的样本代码块的样本代码块特征。

[0103] 步骤506、基于所述样本代码块特征对预设机器学习模型进行训练，生成日志决策模型。

[0104] 步骤507、获取目标代码块。

[0105] 步骤508、从所述目标代码块中提取目标代码块特征。

[0106] 步骤509、将所述目标代码块特征输入至预先训练好的日志决策模型中，根据所述日志决策模型的输出结果判断是否需要在所述目标代码块中插入日志记录语句以及日志级别。

[0107] 步骤510、若根据所述日志决策模型的输出结果确定需要在所述目标代码块中插入日志记录语句，则基于所述日志级别在所述目标代码块中插入日志记录语句。

[0108] 本发明实施例提供的技术方案，可以合理、有效地进行日志记录，减少日志记录的决策时间，不仅能够减少琐碎无效的日志对系统资源的消耗，而且当系统异常时有助于软件开发人员通过合理有效的日志快速找到真正的问题，提高开发效率。

[0109] 需要说明的是，以上各实施例中各技术特征之间的任意排列组合也属于本发明的保护范围。

[0110] 实施例四

[0111] 图6是本发明实施例四提供的一种日志记录装置的示意图，如图6所示，所述装置包括：目标代码块获取模块601、目标代码块特征提取模块602、日志记录语句判断模块603以及日志记录语句插入模块604，其中：

[0112] 目标代码块获取模块601，用于获取目标代码块；

[0113] 目标代码块特征提取模块602，用于从所述目标代码块中提取目标代码块特征；

[0114] 日志记录语句判断模块603，用于将所述目标代码块特征输入至预先训练好的日志决策模型中，根据所述日志决策模型的输出结果判断是否需要在所述目标代码块中插入日志记录语句；

[0115] 日志记录语句插入模块604，用于若根据所述日志决策模型的输出结果判断出需要在所述目标代码块中插入日志记录语句，则在所述目标代码块中插入日志记录语句。

[0116] 本发明实施例通过获取目标代码块，并从目标代码块中提取目标代码块特征，然后将目标代码块特征输入至预先训练好的日志决策模型中，根据日志决策模型的输出结果判断是否需要在目标代码块中插入日志记录语句；若是，则在目标代码块中插入日志记录语句，可以合理、有效地进行日志记录，减少日志记录的决策时间，不仅能够减少琐碎无效的日志对系统资源的消耗，而且当系统异常时有助于软件开发人员通过合理有效的日志快速找到真正的问题，提高开发效率。

[0117] 可选的，所述目标代码块特征提取模块602，用于：

[0118] 将所述目标代码块输入至源码分析工具中，根据所述源码分析工具的输出结果确定目标代码块特征。

[0119] 可选的，所述日志记录语句判断模块603，用于：

[0120] 获取所述日志决策模型的输出分数；

[0121] 若所述输出分数大于预设分数阈值，则确定需要在所述目标代码块中插入日志记录语句；

[0122] 若所述输出分数小于预设分数阈值，则确定不需要在所述目标代码块中插入日志记录语句。

[0123] 可选的，所述装置还包括：

[0124] 样本代码块获取模块，用于在获取目标代码块之前，获取训练项目中的样本代码

块；

[0125] 样本标记模块，用于根据所述样本代码块中是否包含日志记录语句对所述样本代码块进行样本标记；

[0126] 样本代码块特征提取模块，用于提取样本标记后的样本代码块的样本代码块特征；

[0127] 日志决策模型生成模块，用于基于所述样本代码块特征对预设机器学习模型进行训练，生成日志决策模型。

[0128] 可选的，所述装置还包括：

[0129] 预处理模块，用于在基于所述样本代码块特征对预设机器学习模型进行训练之前，对所述样本代码块特征中的文本特征依次进行驼峰转换，小写转换，删除停止词，词干提取和词根化处理以及频率-逆文档频率TF-IDF转换；

[0130] 数值文本特征生成模块，用于基于文本挖掘分类器对经过TF-IDF转换后的文本特征进行降维处理，生成数值文本特征。

[0131] 可选的，所述生成数值文本特征生成模块，用于：

[0132] 基于分层随机抽样将经过TF-IDF转换后的文本特征均分为第一样本和第二样本；

[0133] 基于所述第一样本训练出第一文本挖掘分类器，并基于所述第二样本训练出第二文本挖掘分类器；

[0134] 基于所述第一文本挖掘分类器为所述第一样本分配第一置信度分数，并基于所述第二文本挖掘分类器为所述第二样本分配第二置信度分数；

[0135] 将所述第一置信度分数和所述第二置信度分数作为数值文本特征。

[0136] 可选的，所述装置还包括：

[0137] 日志等级确定模块，用于在根据所述样本代码块中是否包含日志记录语句对所述样本代码块进行样本标记之后，当所述样本代码块中包含日志记录语句时，确定所述样本代码块中包含的日志记录语句的日志等级；

[0138] 日志等级标记模块，用于基于所述日志等级对所述样本代码块进行日志等级标记。

[0139] 可选的，所述日志等级包括致命等级、错误等级、警告等级、信息等级、调试等级和跟踪等级。

[0140] 可选的，所述预设机器学习模型包括随机森林算法、朴素贝叶斯算法、自适应增强Adaboost算法、支持向量机算法。

[0141] 可选的，代码块特征包括文本特征和句法特征；

[0142] 其中，所述文本特征包括代码块的结构特征、代码块调用的方法名称、代码块中声明的变量名称、代码块的类型以及触发策略类型；所述代码块的结构特征包括代码块的源代码行SLOC、代码块调用的方法数目、代码块中声明的变量数目以及代码块中包含日志；

[0143] 所述句法特征包括代码块中是否存在throw语句、是否存在assert语句、是否存在返回值以及是否存在中断异常类型中的任一种。

[0144] 上述日志记录装置可执行本发明任意实施例所提供的日志记录方法，具备执行方法相应的功能模块和有益效果。未在本实施例中详尽描述的技术细节，可参见本发明任意实施例提供的日志记录方法。

[0145] 实施例五

[0146] 本发明实施例五提供了一种基于机器学习的日志决策推荐插件。图7为本发明实施例提供的基于机器学习的日志决策推荐插件的结构图，如图7所示，该日志决策插件包括：代码检测模块、分类器、自动插入日志模块、日志框架配置模块以及动态模板模块。其中，

[0147] 代码检测模块的作用是扫描源码中相应的if和catch代码块，提取代码块的特征值，并将这些特征向量作为分类器模块的输入，并在得到日志决策结果后，如果分类器模块给出的建议是插入日志语句，则会调用自动插入日志模块。可以使用Intellij平台提供的程序结构接口(Program Structure Interfac,PSI)，其负责解析文件，首先构建抽象语法树(Abstract Syntax Tree,AST)定义程序的结构，AST由多个ASTNode组成，每个ASTNode都有一个关联的元素类型，于是针对相应的代码块，只需要获取类型为If或者Catch类型的ASTNode，从而可以获取到所需要的特征值。

[0148] 分类器的作用是为代码检测提供模型支持，其作用是加载前面训练日志推荐算法的模型，在训练出日志决策模型后，将模型文件存储在本地，通过分类器模块进行加载。并对接收到的特征向量进行特征预处理，然后输入到日志决策模型中，得到是否需要插入日志语句以及插入什么等级的日志语句的建议，返回给代码检测模块。

[0149] 自动插入日志模块的作用是接收代码检测模块的调用命令，为开发者提供日志决策的选项，并在开发者选择相应的决策项后，自动插入日志语句。如之前所述，PSI解析源码文件，形成了抽象语法树(AST)，而此模块在AST的树节点上执行操作(例如插入，删除等)，本发明实施例中若需要插入日志记录语句，可由此实现，对AST树结构的修改会立即反映为对基础文档文本的更改。

[0150] 日志框架配置模块的作用为配置项目所需要使用的日志框架，可选log4j、slf4j等。Intellij平台提供了一个API，允许组件或服务，以持续的IDE之间重新启动它们的状态。利用Intellij的组件持久化状态的特性，保存日志框架，也就是相应的日志库文件的标识。其中，在动态模板模块中会使用到此标识。

[0151] 动态模板模块的作用旨在允许开发人员通过关键词插入完整日志语句，其作用是将相应的日志语句和缩写词绑定，并根据配置的log框架信息自动配置和生成logger信息及log语句。此模块通过前面的日志框架配置模块配置的日志框架信息，生成相应的logger对象，以及log语句。

[0152] 本发明实施例中通过构建日志决策模型，确定对开发场景中的if/catch代码块进行日志决策，从而达到辅助开发人员进行日志决策，为其提供日志决策建议的目的。另外，通过日志决策建议插件工具，可以配置日志框架，使用动态模板快速输入日志记录语句，从而帮助开发者更便捷地进行日志框架配置且能更方便地输入日志记录语句。

[0153] 实施例六

[0154] 图8为本发明实施例六提供的一种计算机设备的结构示意图。如图8所示，本申请中的计算机设备可以包括：

[0155] 一个或多个处理器81和存储装置82；该计算机设备的处理器81可以是一个或多个，图8中以一个处理器81为例；存储装置82用于存储一个或多个程序；所述一个或多个程序被所述一个或多个处理器81执行。

[0156] 计算机设备中的处理器81、存储装置82可以通过总线或其他方式连接，图8中以通过总线连接为例。

[0157] 存储装置82作为一种计算机可读存储介质，可设置为存储软件程序、计算机可执行程序以及模块。存储装置82可包括存储程序区和存储数据区，其中，存储程序区可存储操作系统、至少一个功能所需的应用程序；存储数据区可存储根据设备的使用所创建的数据等。此外，存储装置82可以包括高速随机存取存储器，还可以包括非易失性存储器，例如至少一个磁盘存储器件、闪存器件、或其他非易失性固态存储器件。在一些实例中，存储装置82可进一步包括相对于处理器81远程设置的存储器，这些远程存储器可以通过网络连接至计算机设备。上述网络的实例包括但不限于互联网、企业内部网、局域网、移动通信网及其组合。

[0158] 处理器81通过运行存储在存储装置82中的程序，从而执行各种功能应用以及数据处理，例如实现本发明上述实施例所提供的日志记录方法。

[0159] 也即，所述处理单元执行所述程序时实现：获取目标代码块；从所述目标代码块中提取目标代码块特征；将所述目标代码块特征输入至预先训练好的日志决策模型中，根据所述日志决策模型的输出结果判断是否需要在所述目标代码块中插入日志记录语句；若是，则在所述目标代码块中插入日志记录语句。

[0160] 在上述各实施例的基础上，处理器81是设置为通过以下方式从所述目标代码块中提取目标代码块特征：将所述目标代码块输入至源码分析工具中，根据所述源码分析工具的输出结果确定目标代码块特征。

[0161] 在上述各实施例的基础上，处理器81是设置为通过以下方式根据所述日志决策模型的输出结果判断是否需要在所述目标代码块中插入日志记录语句：获取所述日志决策模型的输出分数；若所述输出分数大于预设分数阈值，则确定需要在所述目标代码块中插入日志记录语句；若所述输出分数小于预设分数阈值，则确定不需要在所述目标代码块中插入日志记录语句。

[0162] 在上述各实施例的基础上，在获取目标代码块之前，处理器81还设置为：获取训练项目中的样本代码块；根据所述样本代码块中是否包含日志记录语句对所述样本代码块进行样本标记；提取样本标记后的样本代码块的样本代码块特征；基于所述样本代码块特征对预设机器学习模型进行训练，生成日志决策模型。

[0163] 在上述各实施例的基础上，在基于所述样本代码块特征对预设机器学习模型进行训练之前，处理器81还设置为：对所述样本代码块特征中的文本特征依次进行驼峰转换，小写转换，删除停止词，词干提取和词根化处理以及频率-逆文档频率TF-IDF转换；基于文本挖掘分类器对经过TF-IDF转换后的文本特征进行降维处理，生成数值文本特征。

[0164] 在上述各实施例的基础上，处理器81是设置为通过以下方式基于文本挖掘分类器对经过TF-IDF转换后的文本特征进行降维处理，生成数值文本特征，包括：基于分层随机抽样将经过TF-IDF转换后的文本特征均分为第一样本和第二样本；基于所述第一样本训练出第一文本挖掘分类器，并基于所述第二样本训练出第二文本挖掘分类器；基于所述第一文本挖掘分类器为所述第一样本分配第一置信度分数，并基于所述第二文本挖掘分类器为所述第二样本分配第二置信度分数；将所述第一置信度分数和所述第二置信度分数作为数值文本特征。

[0165] 在上述各实施例的基础上，在根据所述样本代码块中是否包含日志记录语句对所述样本代码块进行样本标记之后，处理器81还设置为：当所述样本代码块中包含日志记录语句时，确定所述样本代码块中包含的日志记录语句的日志等级；基于所述日志等级对所述样本代码块进行日志等级标记。

[0166] 在上述各实施例的基础上，所述日志等级包括致命等级、错误等级、警告等级、信息等级、调试等级和跟踪等级。

[0167] 在上述各实施例的基础上，其特征在于所述预设机器学习模型包括随机森林算法、朴素贝叶斯算法、自适应增强Adaboost算法、支持向量机算法。

[0168] 在上述各实施例的基础上，代码块特征包括文本特征和句法特征；

[0169] 其中，所述文本特征包括代码块的结构特征、代码块调用的方法名称、代码块中声明的变量名称、代码块的类型以及触发策略类型；所述代码块的结构特征包括代码块的源代码行SLOC、代码块调用的方法数目、代码块中声明的变量数目以及代码块中包含日志；

[0170] 所述句法特征包括代码块中是否存在throw语句、是否存在assert语句、是否存在返回值以及是否存在中断异常类型中的任一种。

[0171] 实施例七

[0172] 本发明实施例七还提供一种存储计算机程序的计算机存储介质，所述计算机程序在由计算机处理器执行时用于执行本发明上述实施例任一所述的日志记录方法：获取目标代码块；从所述目标代码块中提取目标代码块特征；将所述目标代码块特征输入至预先训练好的日志决策模型中，根据所述日志决策模型的输出结果判断是否需要在所述目标代码块中插入日志记录语句；若是，则在所述目标代码块中插入日志记录语句。

[0173] 本发明实施例的计算机存储介质，可以采用一个或多个计算机可读的介质的任意组合。计算机可读介质可以是计算机可读信号介质或者计算机可读存储介质。计算机可读存储介质例如可以是——但不限于——电、磁、光、电磁、红外线、或半导体的系统、装置或器件，或者任意以上的组合。计算机可读存储介质的更具体的例子(非穷举的列表)包括：具有一个或多个导线的电连接、便携式计算机磁盘、硬盘、随机存取存储器(RAM)、只读存储器(Read Only Memory，ROM)、可擦式可编程只读存储器((Erasable Programmable Read Only Memory，EPROM)或闪存)、光纤、便携式紧凑磁盘只读存储器(CD-ROM)、光存储器件、磁存储器件、或者上述的任意合适的组合。在本文件中，计算机可读存储介质可以是任何包含或存储程序的有形介质，该程序可以被指令执行系统、装置或者器件使用或者与其结合使用。

[0174] 计算机可读的信号介质可以包括在基带中或者作为载波一部分传播的数据信号，其中承载了计算机可读的程序代码。这种传播的数据信号可以采用多种形式，包括但不限于电磁信号、光信号或上述的任意合适的组合。计算机可读的信号介质还可以是计算机可读存储介质以外的任何计算机可读介质，该计算机可读介质可以发送、传播或者传输用于由指令执行系统、装置或者器件使用或者与其结合使用的程序。

[0175] 计算机可读介质上包含的程序代码可以用任何适当的介质传输，包括——但不限于无线、电线、光缆、射频(Radio Frequency，RF)等等，或者上述的任意合适的组合。

[0176] 可以以一种或多种程序设计语言或其组合来编写用于执行本发明操作的计算机程序代码，所述程序设计语言包括面向对象的程序设计语言—诸如Java、Smalltalk、C++，

还包括常规的过程式程序设计语言——诸如“C”语言或类似的程序设计语言。程序代码可以完全地在用户计算机上执行、部分地在用户计算机上执行、作为一个独立的软件包执行、部分在用户计算机上部分在远程计算机上执行、或者完全在远程计算机或服务器上执行。在涉及远程计算机的情形中，远程计算机可以通过任意种类的网络——包括局域网(LAN)或广域网(WAN)—连接到用户计算机，或者，可以连接到外部计算机(例如利用因特网服务提供商来通过因特网连接)。

[0177] 注意，上述仅为本发明的较佳实施例及所运用技术原理。本领域技术人员会理解，本发明不限于这里所述的特定实施例，对本领域技术人员来说能够进行各种明显的变化、重新调整和替代而不会脱离本发明的保护范围。因此，虽然通过以上实施例对本发明进行了较为详细的说明，但是本发明不仅仅限于以上实施例，在不脱离本发明构思的情况下，还可以包括更多其他等效实施例，而本发明的范围由所附的权利要求范围决定。

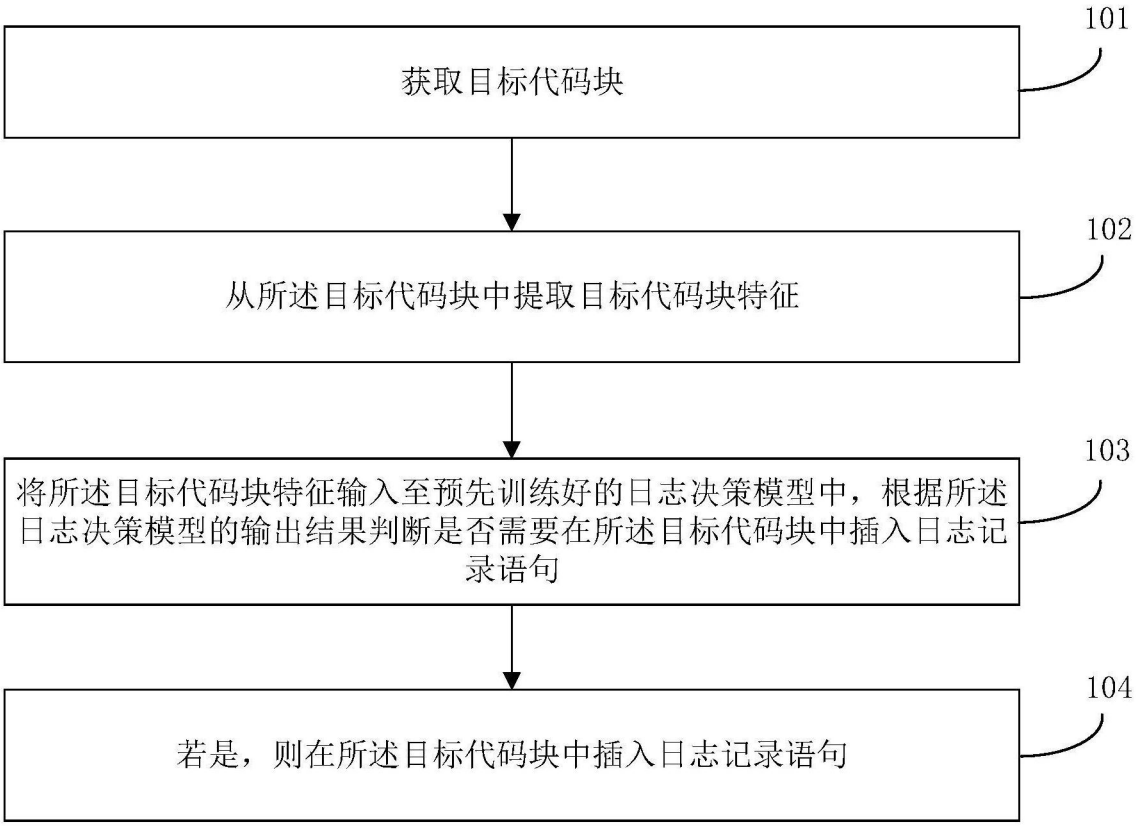

101
获取目标代码块
102
从所述目标代码块中提取目标代码块特征
103
将所述目标代码块特征输入至预先训练好的日志决策模型中，根据所述日志决策模型的输出结果判断是否需要在所述目标代码块中插入日志记录语句
104
若是，则在所述目标代码块中插入日志记录语句

图1

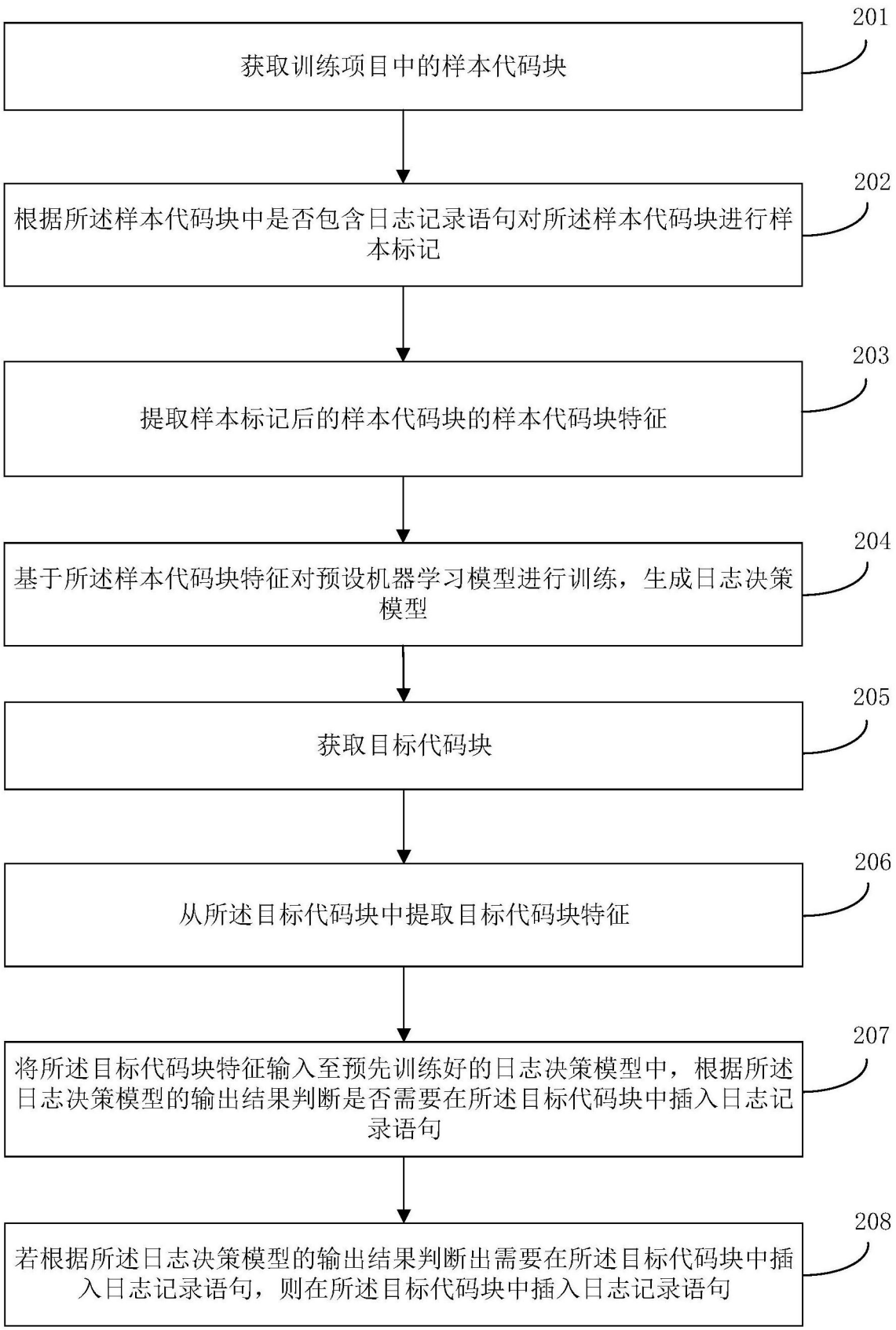

201
获取训练项目中的样本代码块
202
根据所述样本代码块中是否包含日志记录语句对所述样本代码块进行样本标记
203
提取样本标记后的样本代码块的样本代码块特征
204
基于所述样本代码块特征对预设机器学习模型进行训练，生成日志决策模型
205
获取目标代码块
206
从所述目标代码块中提取目标代码块特征
207
将所述目标代码块特征输入至预先训练好的日志决策模型中，根据所述日志决策模型的输出结果判断是否需要在所述目标代码块中插入日志记录语句
208
若根据所述日志决策模型的输出结果判断出需要在所述目标代码块中插入日志记录语句，则在所述目标代码块中插入日志记录语句

图2

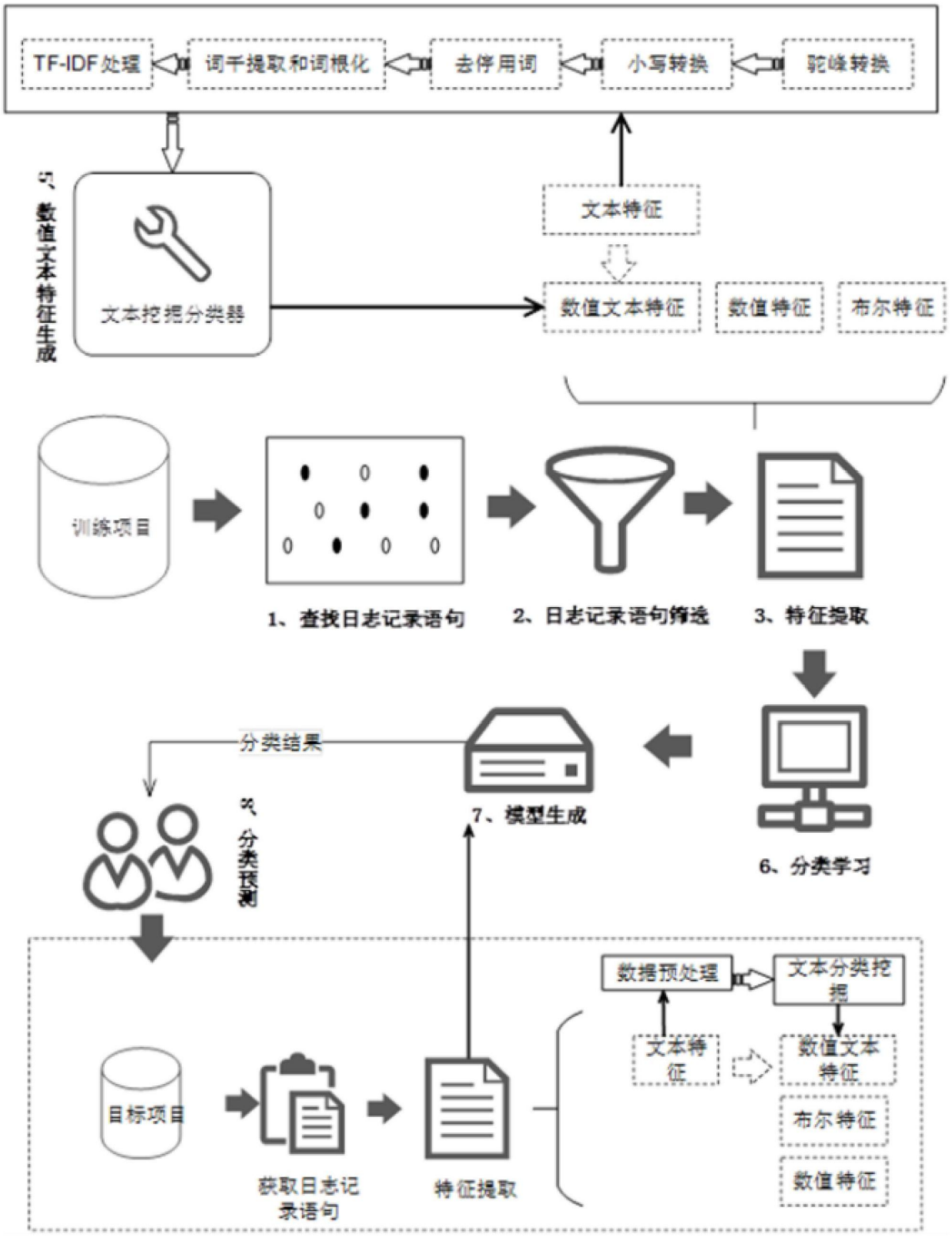

TF-IDF处理
词干提取和词根化
去停用词
小写转换
驼峰转换
5、数值文本特征生成
文本挖掘分类器
文本特征
数值文本特征
数值特征
布尔特征
训练项目
1、查找日志记录语句
2、日志记录语句筛选
3、特征提取
分类结果
4、分类预测
7、模型生成
6、分类学习
数据预处理
文本分类挖掘
文本特征
数值文本特征
布尔特征
数值特征
目标项目
获取日志记录语句
特征提取

图3

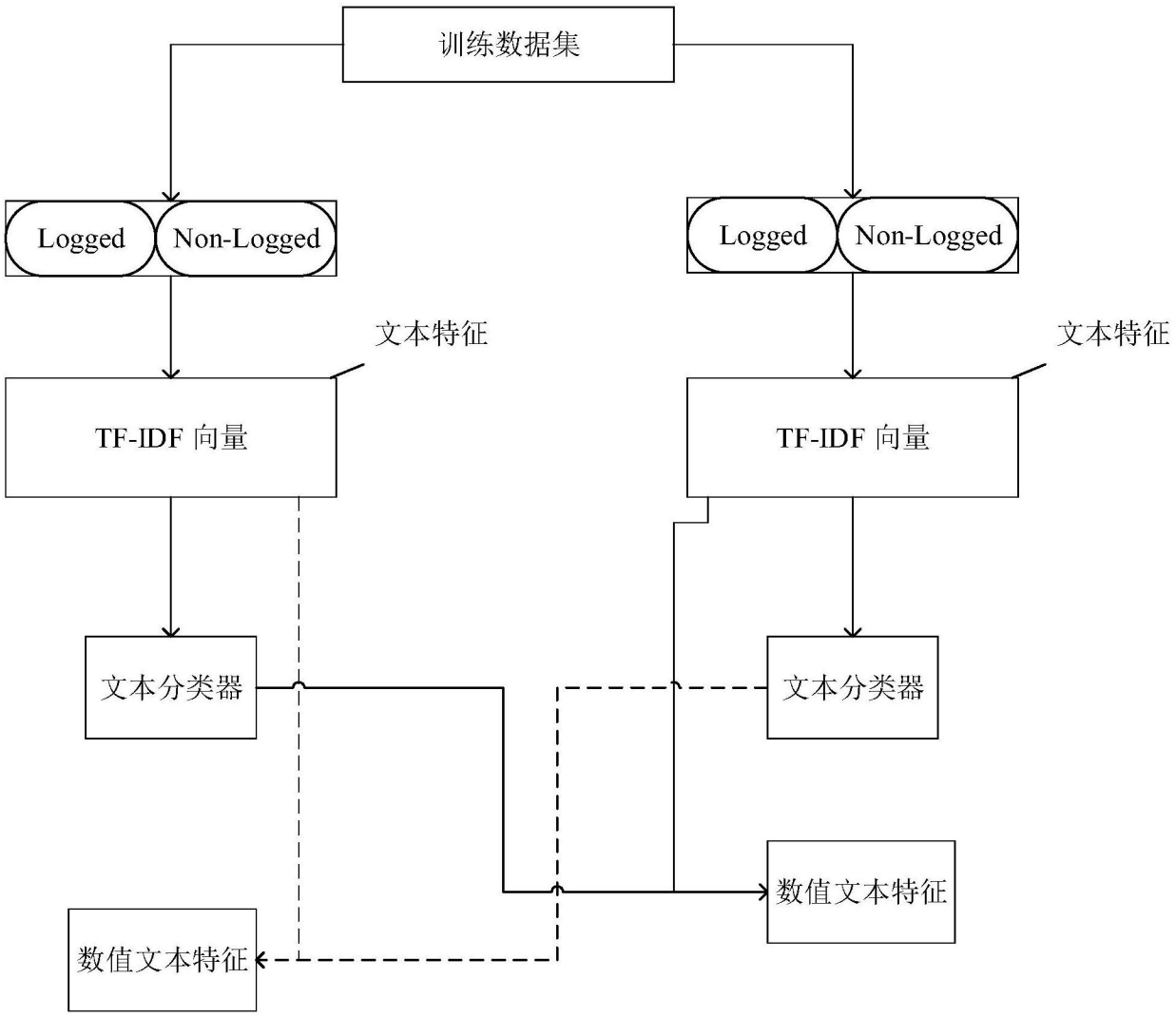

训练数据集
Logged
Non-Logged
Logged
Non-Logged
文本特征
文本特征
TF-IDF 向量
TF-IDF 向量
文本分类器
文本分类器
数值文本特征
数值文本特征

图4

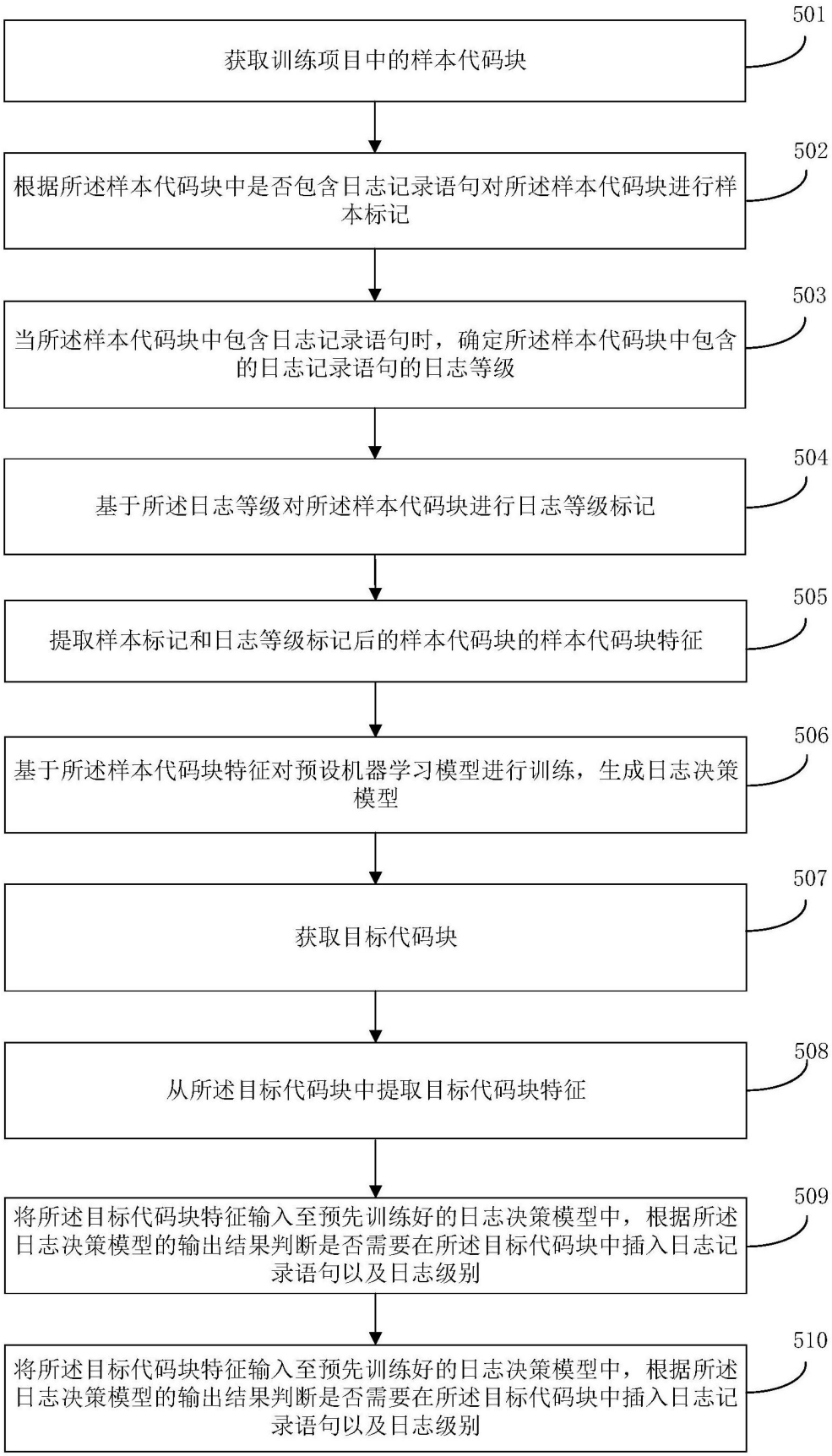

501
获取训练项目中的样本代码块
502
根据所述样本代码块中是否包含日志记录语句对所述样本代码块进行样本标记
503
当所述样本代码块中包含日志记录语句时，确定所述样本代码块中包含的日志记录语句的日志等级
504
基于所述日志等级对所述样本代码块进行日志等级标记
505
提取样本标记和日志等级标记后的样本代码块的样本代码块特征
506
基于所述样本代码块特征对预设机器学习模型进行训练，生成日志决策模型
507
获取目标代码块
508
从所述目标代码块中提取目标代码块特征
509
将所述目标代码块特征输入至预先训练好的日志决策模型中，根据所述日志决策模型的输出结果判断是否需要在所述目标代码块中插入日志记录语句以及日志级别
510
将所述目标代码块特征输入至预先训练好的日志决策模型中，根据所述日志决策模型的输出结果判断是否需要在所述目标代码块中插入日志记录语句以及日志级别

图5

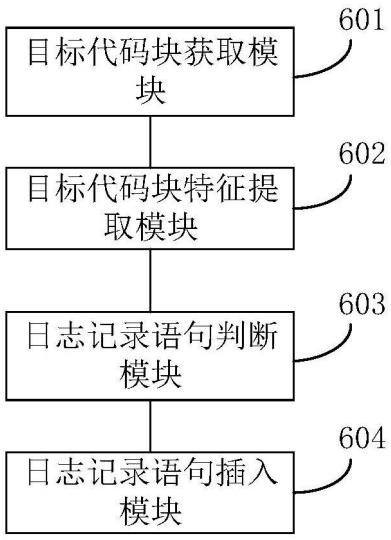

601
目标代码块获取模块
602
目标代码块特征提取模块
603
日志记录语句判断模块
604
日志记录语句插入模块

图6

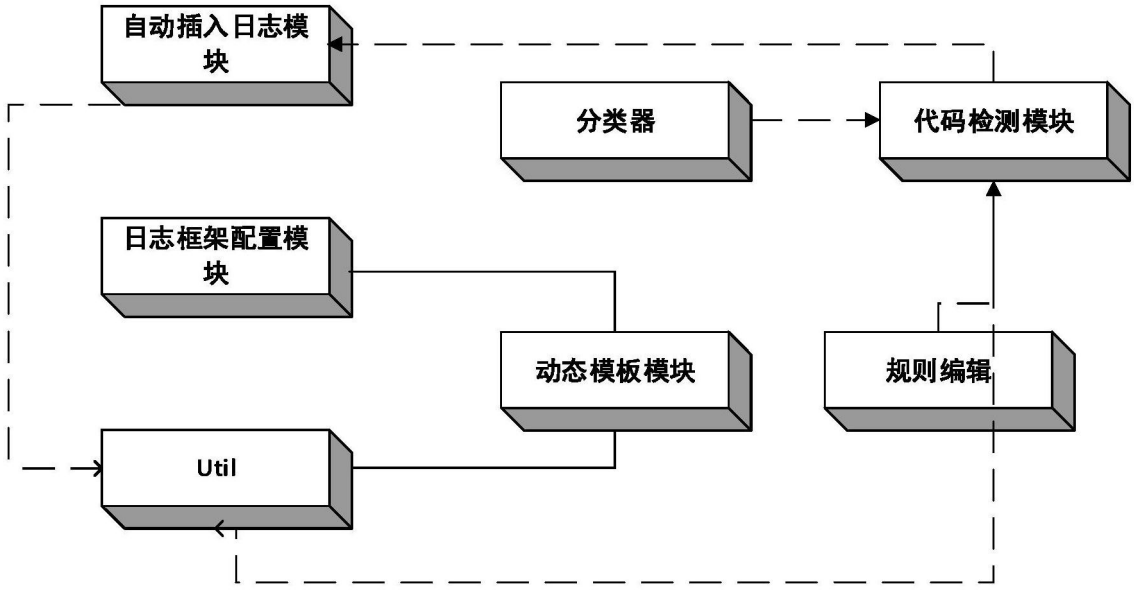

自动插入日志模块
分类器
代码检测模块
日志框架配置模块
动态模板模块
规则编辑
Util

图7

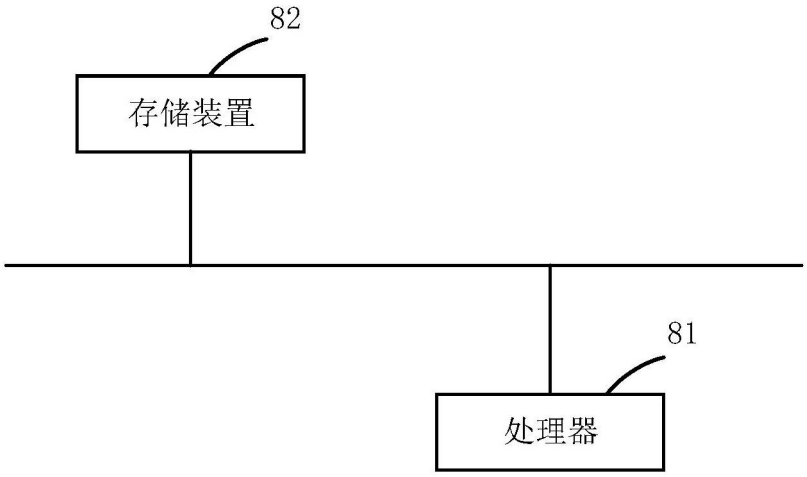



图8